

# Data preparation

## Dataset: 2020 NFL Predictions

### Comparing data types between sheet 1 and sheet 2

##	Column	Sheet1_Type	Sheet2_Type
##	playoff	playoff	logical
##	elo1_pre	elo1_pre	character
##	elo2_post	elo2_post	numeric
##	qb1_value_pre	qb1_value_pre	numeric
##	score1	score1	character
##	score2	score2	character

### Convert character data type to numeric data type and combine the sheets

```
## tibble [269 × 30] (S3: tbl_df/tbl/data.frame)
## $ date      : num [1:269] 44084 44087 44087 44087 44087 ...
## $ season    : num [1:269] 2020 2020 2020 2020 2020 2020 2020 2020 2 2020 2020 ...
## $ neutral   : num [1:269] 0 0 0 0 0 0 0 0 0 99 ...
## $ playoff   : chr [1:269] NA NA NA NA ...
## $ team1     : chr [1:269] "KC" "MIN" "BUF" "ATL" ...
## $ team2     : chr [1:269] "Houston" "GB" "NYJ" "SEA" ...
## $ elo1_pre  : num [1:269] 1665 1571 1511 1535 1439 ...
## $ elo2_pre  : num [1:269] 1528 1582 1458 1547 1483 ...
## $ elo_prob1 : num [1:269] 0.762 0.577 0.664 0.575 0.53 ...
## $ elo_prob2 : num [1:269] 0.238 0.423 0.336 0.425 0.47 ...
## $ elo1_post : num [1:269] 1677 1544 1527 1503 1458 ...
## $ elo2_post : num [1:269] 1516 1610 1443 1578 1463 ...
## $ qbelo1_pre : num [1:269] 1651 1545 1533 1501 1392 ...
## $ qbelo2_pre : num [1:269] 1497 1555 1451 1544 1518 ...
## $ qb1       : chr [1:269] "Patrick Mahomes" "Kirk Cousins" "Josh Allen" "Matt
Ryan" ...
## $ qb2       : chr [1:269] "Deshaun Watson" "Aaron Rodgers" "Sam Darnold" "Rus
sell Wilson" ...
## $ qb1_value_pre : num [1:269] 240 158 153 180 122 ...
## $ qb2_value_pre : chr [1:269] "195.61581086259901" "176.74031012735901" "134.1209
53889665" "193.189148644591" ...
## $ qb1_adj    : chr [1:269] "6.9428085552971899" "-0.81276266285677301" "0.6006
1032740741205" "-1.73926316029488" ...
## $ qb2_adj    : num [1:269] 3.63 0.49 5.19 4.24 6.98 ...
## $ qbelo_prob1 : num [1:269] 0.752 0.531 0.655 0.49 0.352 ...
## $ qbelo_prob2 : num [1:269] 0.248 0.469 0.345 0.51 0.648 ...
## $ qb1_game_value: num [1:269] 251 273 381 322 217 ...
## $ qb2_game_value: num [1:269] 163.1 487.2 93.6 440.2 267 ...
## $ qb1_value_post: num [1:269] 241 170 176 194 131 ...
## $ qb2_value_post: num [1:269] 192 208 130 218 167 ...
## $ qbelo1_post  : num [1:269] 1664 1520 1549 1475 1420 ...
## $ qbelo2_post  : num [1:269] 1485 1580 1435 1569 1490 ...
## $ score1      : num [1:269] 34 34 27 25 27 38 21 30 23 27 ...
## $ score2      : num [1:269] 20 43 17 38 20 6 11 34 27 17 ...
```

### 1. The dates are in numeric format, let's convert them to date format.

```
##
## 1905-07-12 2020-09-10 2020-09-13 2020-09-14 2020-09-17 2020-09-20 2020-09-21
##          1          1          13          2          1          14          1
## 2020-09-24 2020-09-27 2020-09-28 2020-10-01 2020-10-04 2020-10-05 2020-10-08
##          1          14          1          1          12          2          1
## 2020-10-11 2020-10-12 2020-10-13 2020-10-18 2020-10-19 2020-10-22 2020-10-25
##          11          1          1          12          2          1          11
## 2020-10-26 2020-10-29 2020-11-01 2020-11-02 2020-11-05 2020-11-08 2020-11-09
##          1          1          12          1          1          12          1
## 2020-11-12 2020-11-15 2020-11-16 2020-11-19 2020-11-22 2020-11-23 2020-11-26
##          1          12          1          1          12          1          2
## 2020-11-29 2020-11-30 2020-12-02 2020-12-06 2020-12-07 2020-12-08 2020-12-10
##          12          1          1          12          2          1          1
## 2020-12-13 2020-12-14 2020-12-17 2020-12-19 2020-12-20 2020-12-21 2020-12-25
##          14          1          1          2          12          1          1
## 2020-12-26 2020-12-27 2020-12-28 2021-01-03 2021-01-09 2021-01-10 2021-01-16
##          3          11          1          16          3          3          2
## 2021-01-17 2021-01-24 2021-02-07
##          2          2          1
```

## 2. Removing a row with an unusual date (1905-07-12) as it is irrelevant in a 2020 NFL prediction dataset

```
##
## 2020-09-10 2020-09-13 2020-09-14 2020-09-17 2020-09-20 2020-09-21 2020-09-24
##          1          13          2          1          14          1          1
## 2020-09-27 2020-09-28 2020-10-01 2020-10-04 2020-10-05 2020-10-08 2020-10-11
##          14          1          1          12          2          1          11
## 2020-10-12 2020-10-13 2020-10-18 2020-10-19 2020-10-22 2020-10-25 2020-10-26
##          1          1          12          2          1          11          1
## 2020-10-29 2020-11-01 2020-11-02 2020-11-05 2020-11-08 2020-11-09 2020-11-12
##          1          12          1          1          12          1          1
## 2020-11-15 2020-11-16 2020-11-19 2020-11-22 2020-11-23 2020-11-26 2020-11-29
##          12          1          1          12          1          2          12
## 2020-11-30 2020-12-02 2020-12-06 2020-12-07 2020-12-08 2020-12-10 2020-12-13
##          1          1          12          2          1          1          14
## 2020-12-14 2020-12-17 2020-12-19 2020-12-20 2020-12-21 2020-12-25 2020-12-26
##          1          1          2          12          1          1          3
## 2020-12-27 2020-12-28 2021-01-03 2021-01-09 2021-01-10 2021-01-16 2021-01-17
##          11          1          16          3          3          2          2
## 2021-01-24 2021-02-07
##          2          1
```

## 3. All values of 'season' should be 2020 since this prediction was done on a 2020 dataset.

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2     2020     2020    1998     2020     2020
```

Not all values are 2020. Lets correct them.

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2020     2020     2020    2020     2020    2020
```

#### 4. Handling missing values in 'playoff' column and removing the 'neutral' column

Labeling them as NA so it is easy to analyze results Also, let's flag and remove the 'neutral' column because except for 3 errored values (9,99,99) and others are 0, meaning they home/away games.

```
## [1] "Summary of playoff\n"
```

```
##
##      c      d      s      w <NA>
##      2      4      1      6  255
```

#### 5. Handling missing values in the 'team 1' column

```
##
##  ARI  ATL  BAL  BUF  CAR  CHI  CIN  CLE  DAL  DEN  DET  GB  HOU  IND  JAX  KC
##    7    8    8   10    8    8    8    8    8    8    8    9    7    8    8   10
##  LAC  LAR  MIA  MIN  NE   NO   NYG  NYJ  OAK  PHI  PIT  SEA  SF   TB   TEN  WSH
##    8    7    8    8    8   10    8    8    8    8    9    9    8    8    9    9
## <NA>
##    4
```

```
## # A tibble: 4 × 2
##   team1 qb1
##   <chr> <chr>
## 1 <NA>  Kyler Murray
## 2 <NA>  Aaron Rodgers
## 3 <NA>  John Wolford
## 4 <NA>  <NA>
```

There are 4 rows with NA. Let's remove the row where all values are empty. Let's find and impute the team names from the quarterbacks respective to these rows.

```
##
##  ARI  ATL  BAL  BUF  CAR  CHI  CIN  CLE  DAL  DEN  DET  GB  HOU  IND  JAX  KC  LAC  LAR  MIA  MIN
##    8    8    8   10    8    8    8    8    8    8    8   10    7    8    8   10    8    8    8    8
##  NE   NO  NYG  NYJ  OAK  PHI  PIT  SEA   SF   TB  TEN  WSH
##    8   10    8    8    8    8    9    9    8    8    9    9
```

#### 6. Handling missing values in the 'team 2' column

```
##
##      ARI      ATL      BAL      BUF      CAR      CHI      CIN      CLE      DAL      DEN
##      8        8       10        9        8        8        8       10        8        8
##      DET      GB      HOU Houston      IND      JAX      KC      LAC      LAR      MIA
##      8        7        7        1        9        8        8        8       10        7
##      MIN      NE      NO      NYG      NYJ      OAK OAKLAND      PHI      PIT      SEA
##      8        8        8        8        8        7        1        7        8        8
##      SF      TB      TEN      WSH      <NA>
##      8       11        8        8        3
```

```
## # A tibble: 3 × 2
##   team2 qb2
##   <chr> <chr>
## 1 <NA> Mitchell Trubisky
## 2 <NA> Carson Wentz
## 3 <NA> Tua Tagovailoa
```

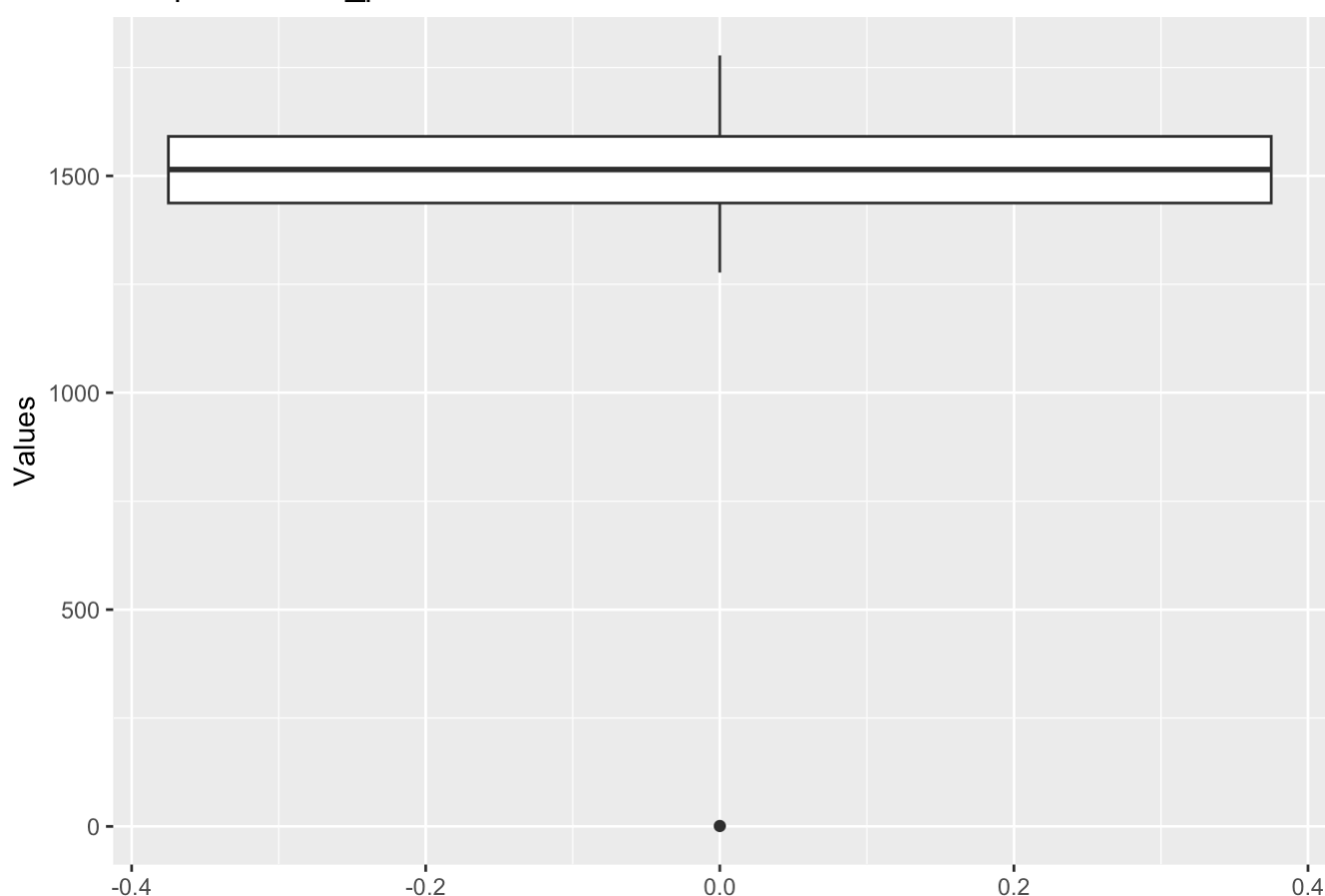
1 - Three missing values, let's deal with them as we did for 'team 1' column 2 - Abbreviations of Houston, and oakland has to be corrected

```
##
## ARI ATL BAL BUF CAR CHI CIN CLE DAL DEN DET GB HOU IND JAX KC LAC LAR MIA MIN
## 8 8 10 9 8 9 8 10 8 8 8 7 8 9 8 8 8 10 8 8
## NE NO NYG NYJ OAK PHI PIT SEA SF TB TEN WSH
## 8 8 8 8 8 8 8 8 8 11 8 8
```

**7. elo1\_pre: The Elo rating for team 1 before the match. Let's handle outliers and missing values**

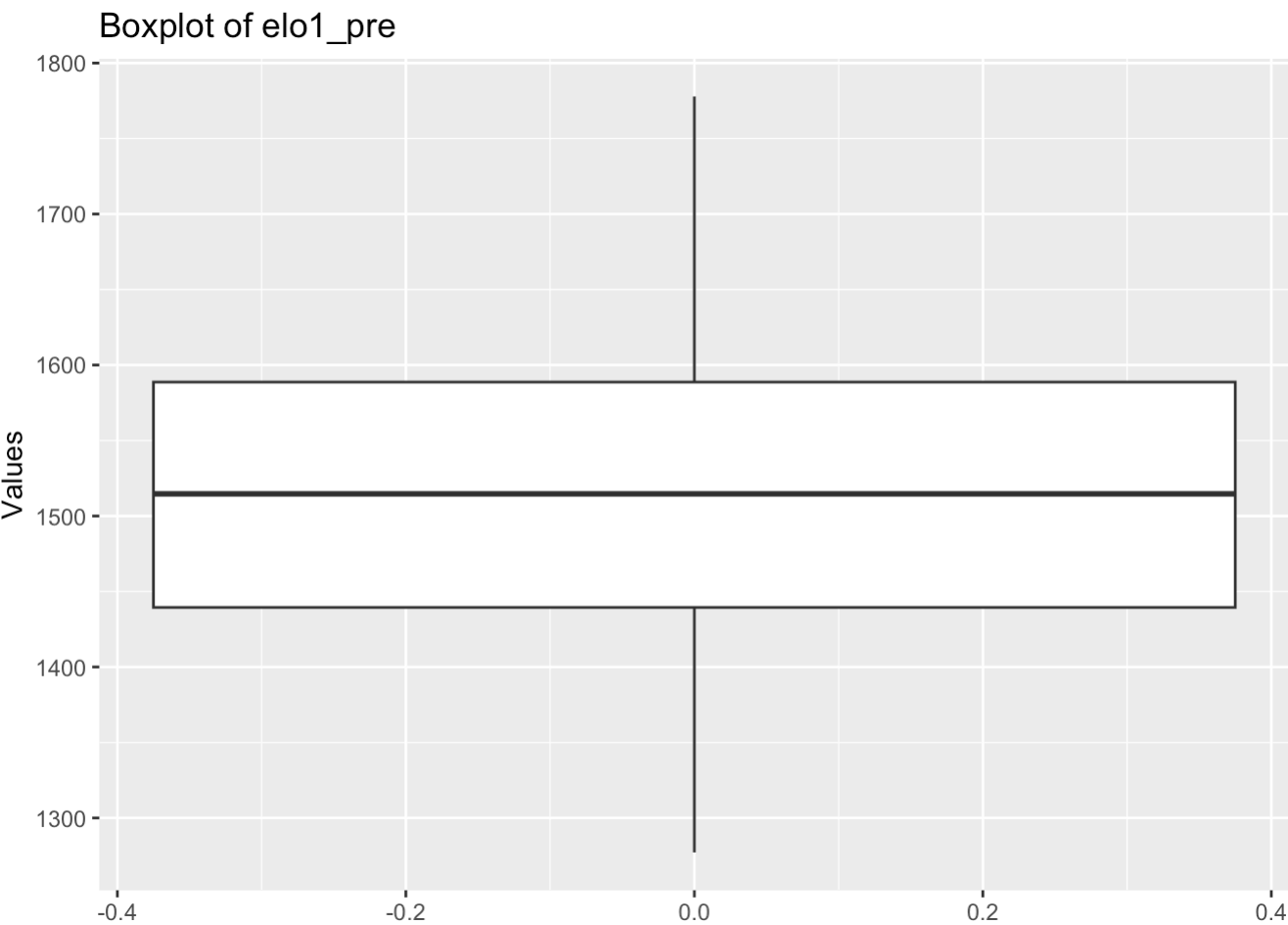
```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##     1    1437    1515    1510    1591    1778     4
```

Boxplot of elo1\_pre



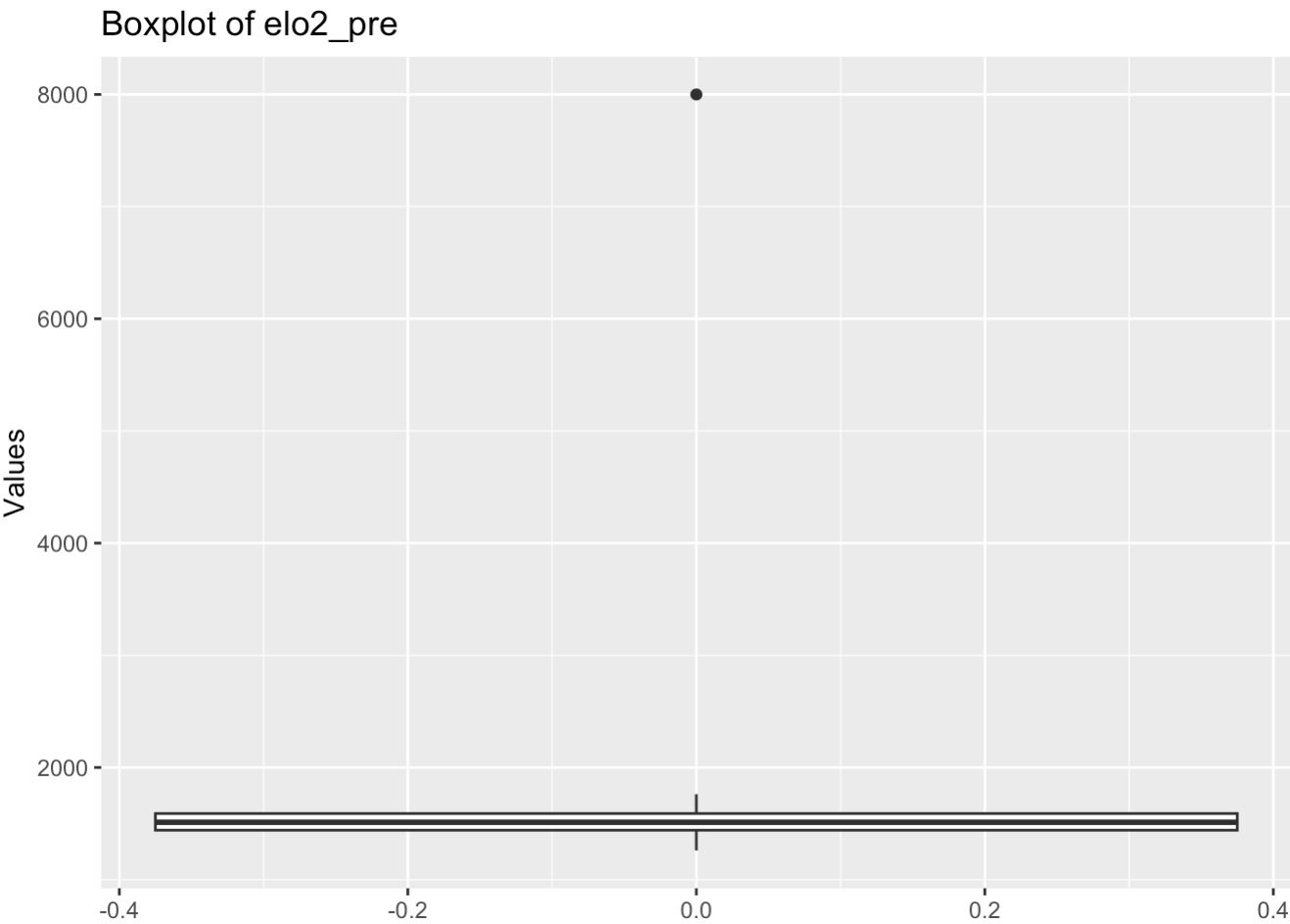
50% of the data falls within a range close to 1500 while the rest ranging from around 1200 to 1800. There is one outlier around 0, represented by the dot. Let's impute the missing values and the outlier with the median value. Since we have outliers, it is best to impute with median instead of mean

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	1277	1439	1515	1515	1589	1778



8. elo2\_pre: The Elo rating for team 2 before the match. Doing the same as we did for elo1\_pre

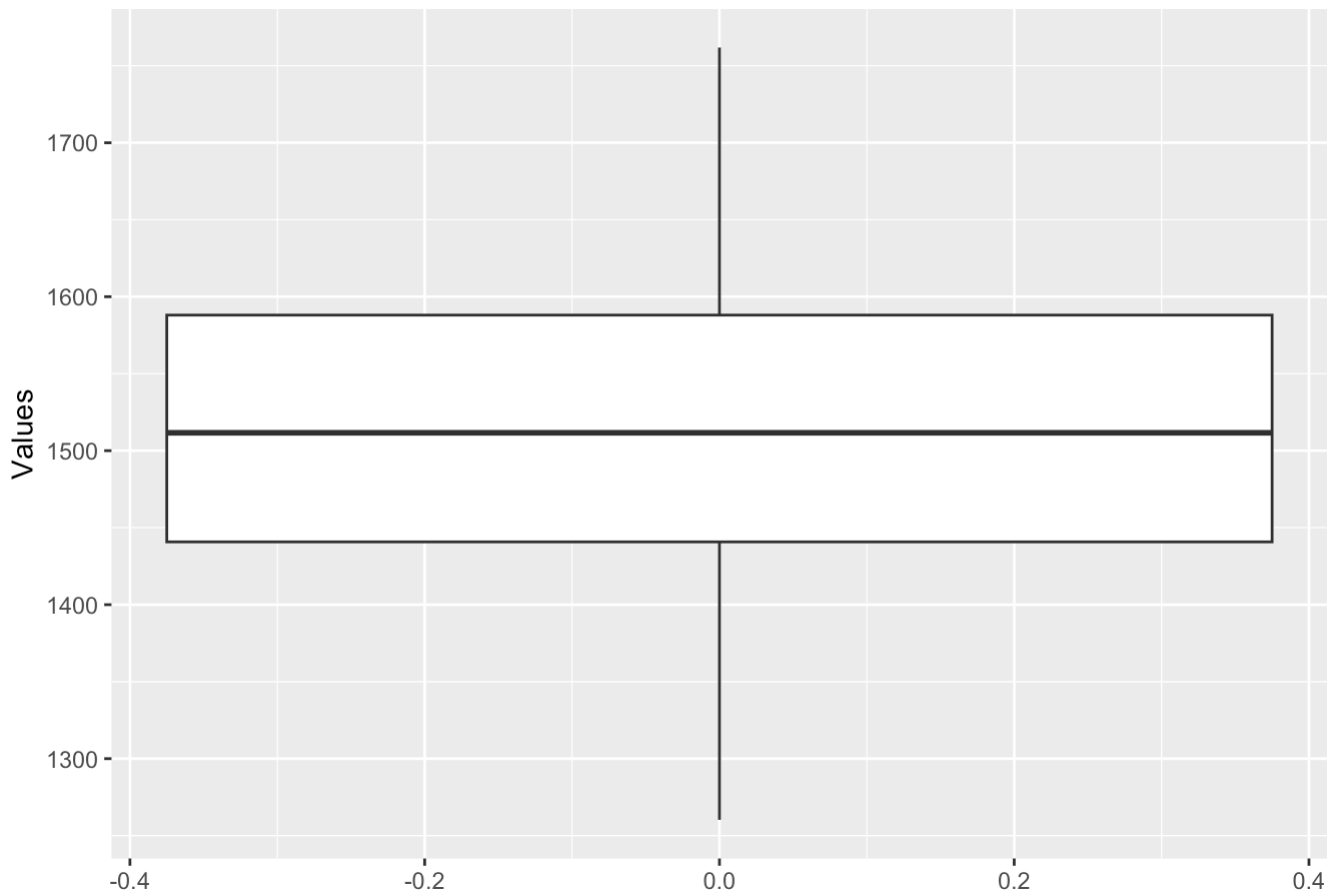
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	1260	1441	1512	1533	1589	8000	2



Its the opposite with this column, there is an outlier around 8000 and 2 missing values.

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	1260	1441	1512	1509	1588	1762

Boxplot of elo2\_pre



## 9. Handling outliers and missing values in columns elo\_prob1, and elo\_prob2

```
##
## Summary for elo_prob1 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##  0.1720  0.4665  0.5993  0.5869  0.7071  0.9370     2
##
## Summary for elo_prob2 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##   0.063   0.293   0.401  15.506   0.534 4000.000     2
```

Upper bound outlier found in elo\_prob2 (Max - 4000). Replace values of elo\_prob2 with  $(1 - \text{elo\_prob1})$  according to the formula

```
##
## Summary for elo_prob1 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1720  0.4669  0.5993  0.5870  0.7067  0.9370
##
## Summary for elo_prob2 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.06298 0.29325 0.40065 0.41302 0.53313 0.82805
```

## 10. Handling negative outliers and missing values in columns elo1\_post, and elo2\_post

```
##
## Summary for elo1_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1    1430    1514    1497    1586    1778         4
##
## Summary for elo2_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -1504    1435    1510    1500    1591    1775         7
```

elo1\_post - outlier at 0, it should be the value 1 (min) and two missing values. elo2\_post - negative outlier, elo ratings are never negative. Let's impute it with median

```
##
## Summary for elo1_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1260    1438    1514    1509    1585    1778
##
## Summary for elo2_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1256    1437    1510    1512    1590    1775
```

## 11. Handling outliers and missing values in columns qbelo1\_pre, and qbelo2\_pre

```
##
## Summary for qbelo1_pre :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0    1438    1514    1497    1580    1757         4
##
## Summary for qbelo2_pre :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1259    1441    1516    1512    1581    1742         5
```

qbelo1\_pre - Quarterback Elo ratings, a value of 0 is highly unlikely and unrealistic. It looks like an error, so, lets impute them with median qbelo2\_pre - Missing values

```
##
## Summary for qbelo1_pre :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1272    1446    1514    1514    1579    1757
##
## Summary for qbelo2_pre :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1259    1442    1516    1512    1579    1742
```

## 12. Handling missing values in columns qb1, and qb2

```
## # A tibble: 4 × 2
##   qb1    team1
##   <chr> <chr>
## 1 <NA>    TB
## 2 <NA>    CAR
## 3 <NA>    JAX
## 4 <NA>    KC
```



```
## # A tibble: 1 × 2
##   qb2    team2
##   <chr> <chr>
## 1 <NA>   IND
```

Let's find and update the quarterback names from the teams we found and remove the row where all values are NA. Also, scan for spelling mistakes or special characters. None found

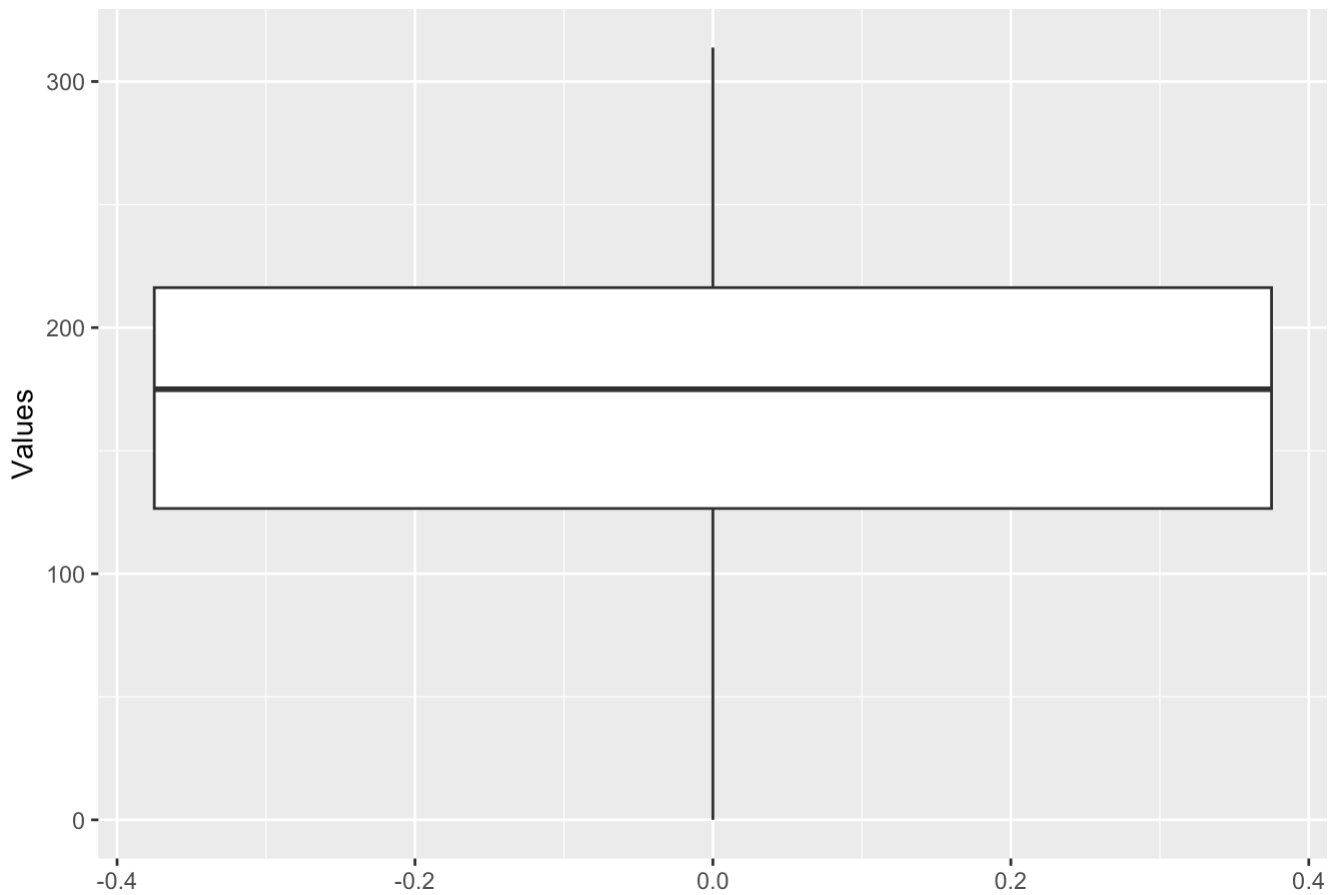
```
##
##   Aaron Rodgers      Alex Smith      Andy Dalton      Baker Mayfield
##           10              1              4              8
## Ben Roethlisberger  Brandon Allen    C.J. Beathard      Cam Newton
##           9              3              1              8
##   Carson Wentz      Chad Henne      Colt McCoy      Dak Prescott
##           6              1              1              3
##   Daniel Jones      Derek Carr      Deshaun Watson      Drew Brees
##           7              8              7              9
##   Drew Lock      Dwayne Haskins  Gardner Minshew      Garrett Gilbert
##           6              4              4              1
##   Jake Luton      Jalen Hurts      Jared Goff      Jeff Driskel
##           1              2              7              1
##   Jimmy Garoppolo    Joe Burrow      Joe Flacco      John Wolford
##           3              4              2              1
##   Josh Allen      Justin Herbert    Kendall Hinton      Kirk Cousins
##          10              8              1              8
##   Kyle Allen      Kyler Murray      Lamar Jackson      Matt Ryan
##           3              8              8              8
##   Matthew Stafford  Mike Glennon    Mitchell Trubisky      Nick Foles
##           8              3              4              4
##   Nick Mullens      P.J. Walker      Patrick Mahomes      Philip Rivers
##           4              1              9              8
##   Russell Wilson      Ryan Finley      Ryan Fitzpatrick      Ryan Tannehill
##           9              1              3              9
##   Sam Darnold      Taylor Heinicke    Taysom Hill      Teddy Bridgewater
##           6              1              1              7
##   Tom Brady      Tua Tagovailoa
##           8              5
```

```
##
##      Aaron Rodgers      Alex Smith      Andy Dalton      Baker Mayfield
##              7              5              5              10
##      Ben DiNucci Ben Roethlisberger      Brandon Allen      Brett Rypien
##              1              7              2              1
##      Brian Hoyer      C.J. Beathard      Cam Newton      Carson Wentz
##              1              1              7              6
##      Colt McCoy      Dak Prescott      Daniel Jones      Derek Carr
##              1              2              7              8
##      Deshaun Watson      Drew Brees      Drew Lock      Dwayne Haskins
##              8              5              7              2
##      Gardner Minshew      Jake Luton      Jalen Hurts      Jared Goff
##              5              1              2              9
##      Jimmy Garoppolo      Joe Burrow      Joe Flacco      John Wolford
##              3              6              2              1
##      Josh Allen      Justin H      Justin Herbert      Kirk Cousins
##              9              1              6              8
##      Kyle Allen      Kyler Murray      Lamar Jackson      Mason Rudolph
##              1              8              9              1
##      Matt Ryan      Matthew Stafford      Mike Glennon      Mitchell Trubisky
##              8              8              2              6
##      Nick Foles      Nick Mullens      Patrick Mahomes      Philip Rivers
##              3              4              8              9
##      Robert Griffin III      Russell Wilson      Ryan Fitzpatrick      Ryan Tannehill
##              1              8              4              8
##      Sam Darnold      Taysom Hill      Teddy Bridgewater      Tom Brady
##              6              3              8              11
##      Tua Tagovailoa      Tyrod Taylor
##              4              1
```

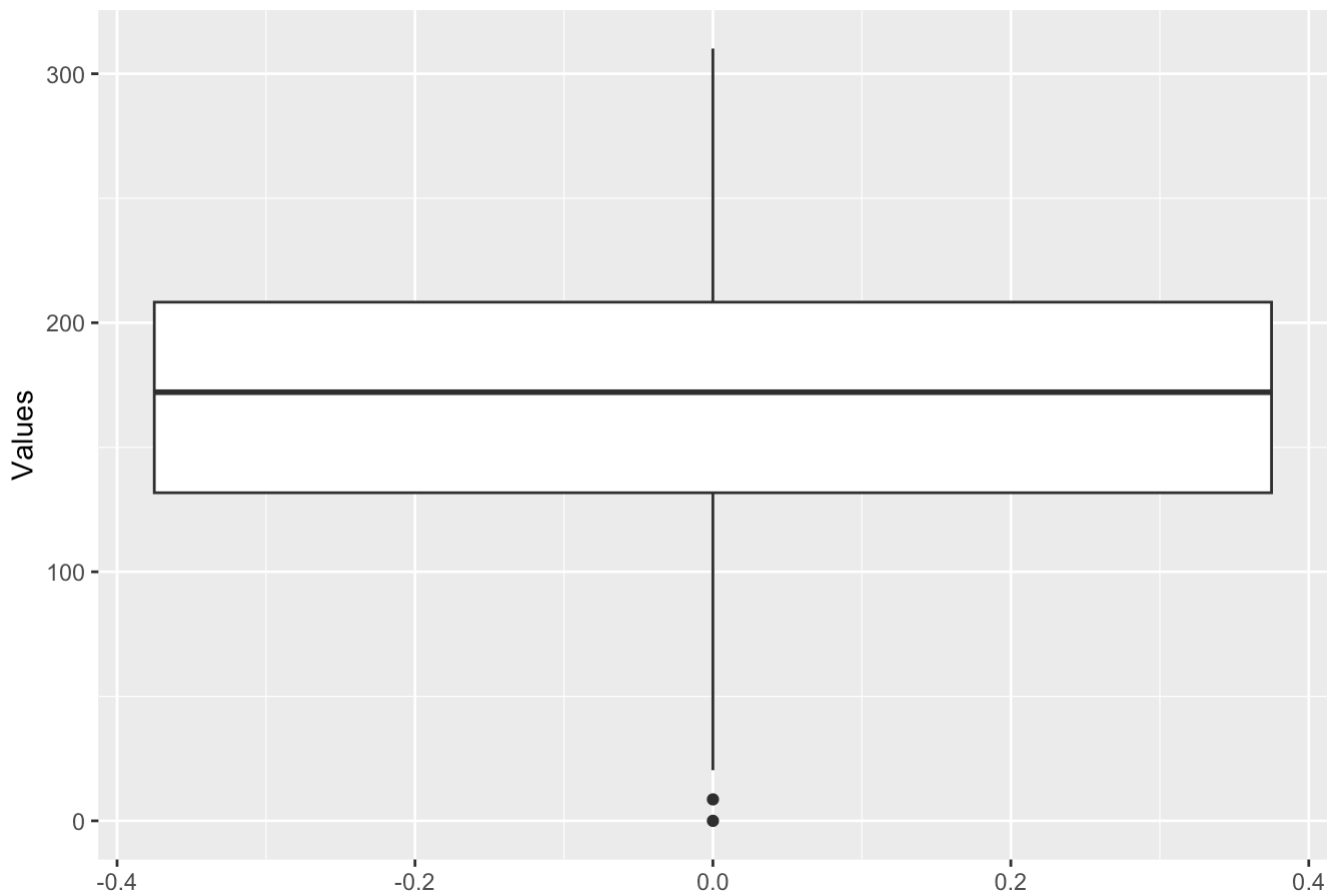
### 13. Handling missing values in columns qb1\_value\_pre, and qb2\_value\_pre

```
##
## Summary for qb1_value_pre :
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      0.0   126.5   175.0   170.6   216.2   313.8      4
##
## Summary for qb2_value_pre :
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      0.0   131.7   172.1   169.8   208.3   310.1      5
```

Boxplot of qb1\_value\_pre



Boxplot of qb2\_value\_pre



potential outliers around 0, a realistic minimum value for a quarterback's Elo rating in established leagues should be above 100 or 150, as Elo ratings typically start around 1500.

```
##
## Summary for qb1_value_pre :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  102.2   150.2   175.0   184.5   216.1   313.8
##
## Summary for qb2_value_pre :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   102.0   145.7   172.1   180.2   207.8   310.1
```

#### 14. Handling missing values in columns qb1\_adj, and qb2\_adj

```
##
## Summary for qb1_adj :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -242.488  -6.634    6.350   -1.731   15.692   54.827      8
##
## Summary for qb2_adj :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -218.569  -4.025    6.120   -1.464   16.223   53.096      3
```

Both the adjusted values have outliers and NAs. let's find the lower and upper bounds by calculating IQR. And, replace the outliers < lower bound | outliers > upper bound | NAs with median

```
##
## Summary for qb1_adj :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -30.4603  -0.8247    6.3501    7.6494   15.3847   48.9537
##
## Summary for qb2_adj :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -34.3734  -0.3922    6.1197    8.0197   15.7942   45.8382
```

#### 15. Handling missing values in columns qbelo\_prob1, and qbelo\_prob2

```
##
## Summary for qbelo_prob1 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.07023  0.40905  0.54902  0.64160  0.69645  25.00000      3
##
## Summary for qbelo_prob2 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.06693  0.30426  0.45370  0.45142  0.59095  0.92977      3
```

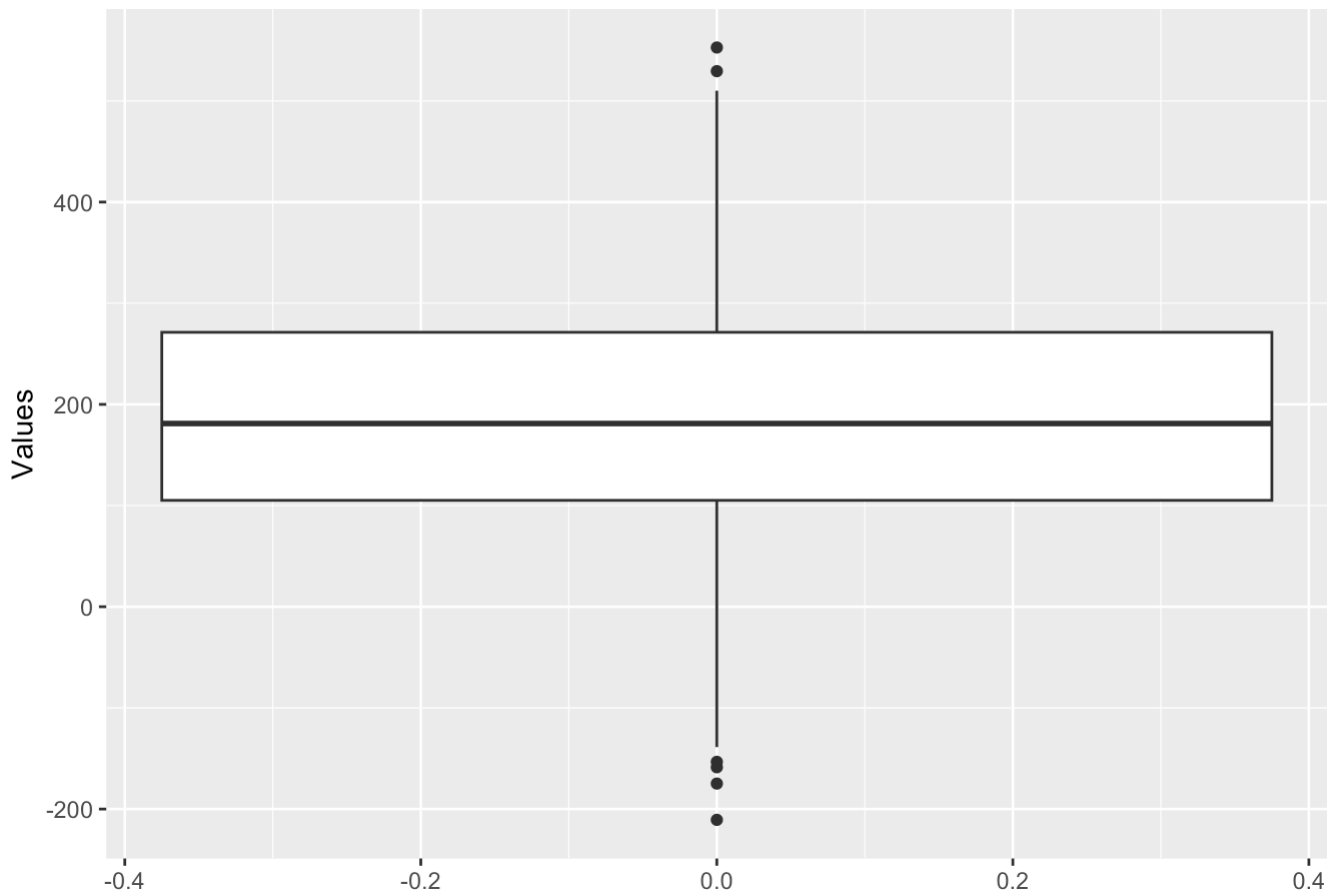
Both have 3 NAs each, qbelo\_prob1 alone has an outlier (25)

```
##
## Summary for qbelo_prob1 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.07023  0.40969  0.54902  0.54898  0.69531  0.93307
##
## Summary for qbelo_prob2 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.06693  0.30469  0.45370  0.45144  0.59031  0.92977
```

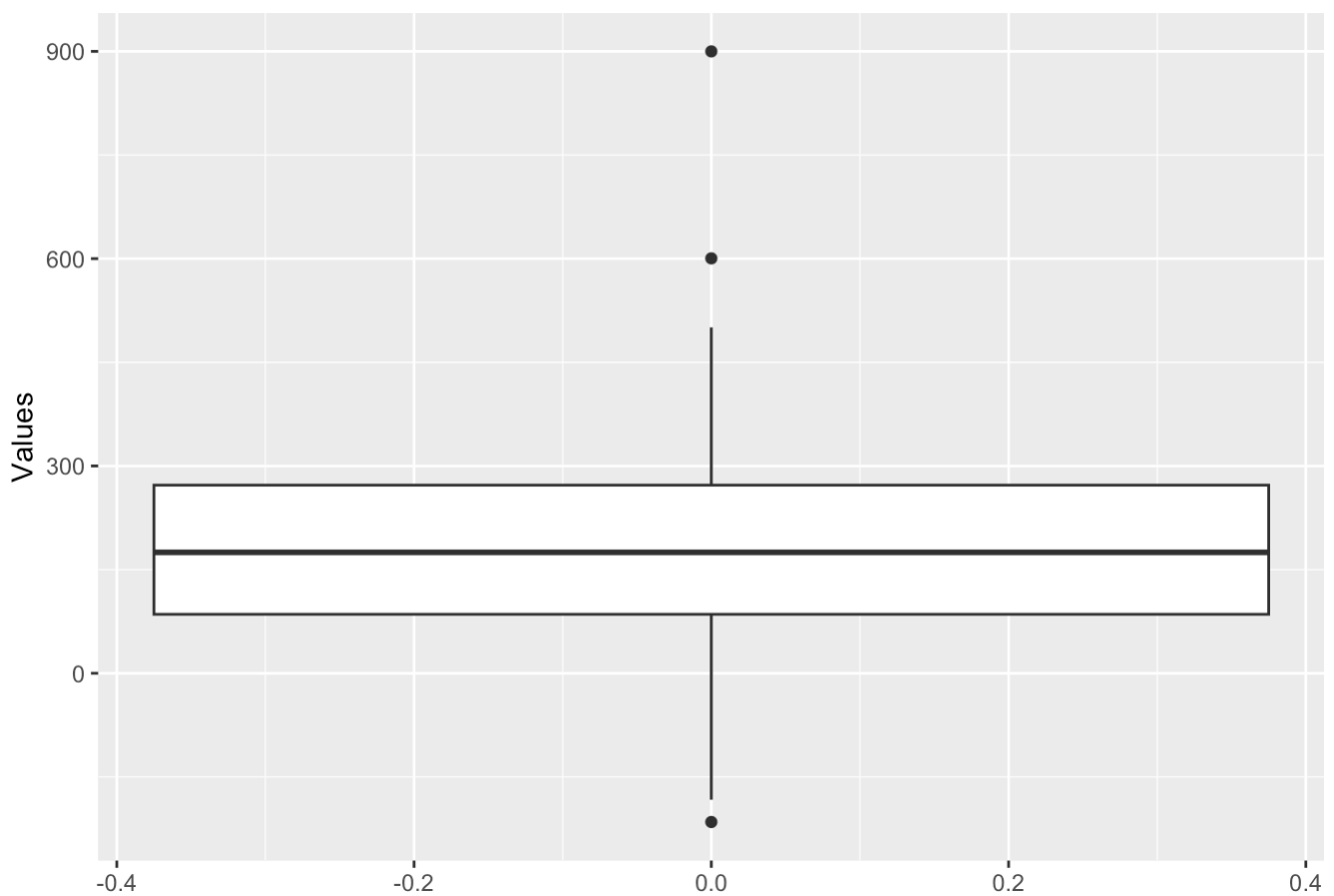
## 16. Handling missing values in columns qb1\_game\_value, and qb2\_game\_value

```
##  
## Summary for qb1_game_value :  
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
## -210.7   105.2   181.1   181.5   271.3   552.8     5  
##  
## Summary for qb2_game_value :  
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
## -215.31   85.33  174.96  179.26  272.24  900.00     5
```

Boxplot of qb1\_game\_value



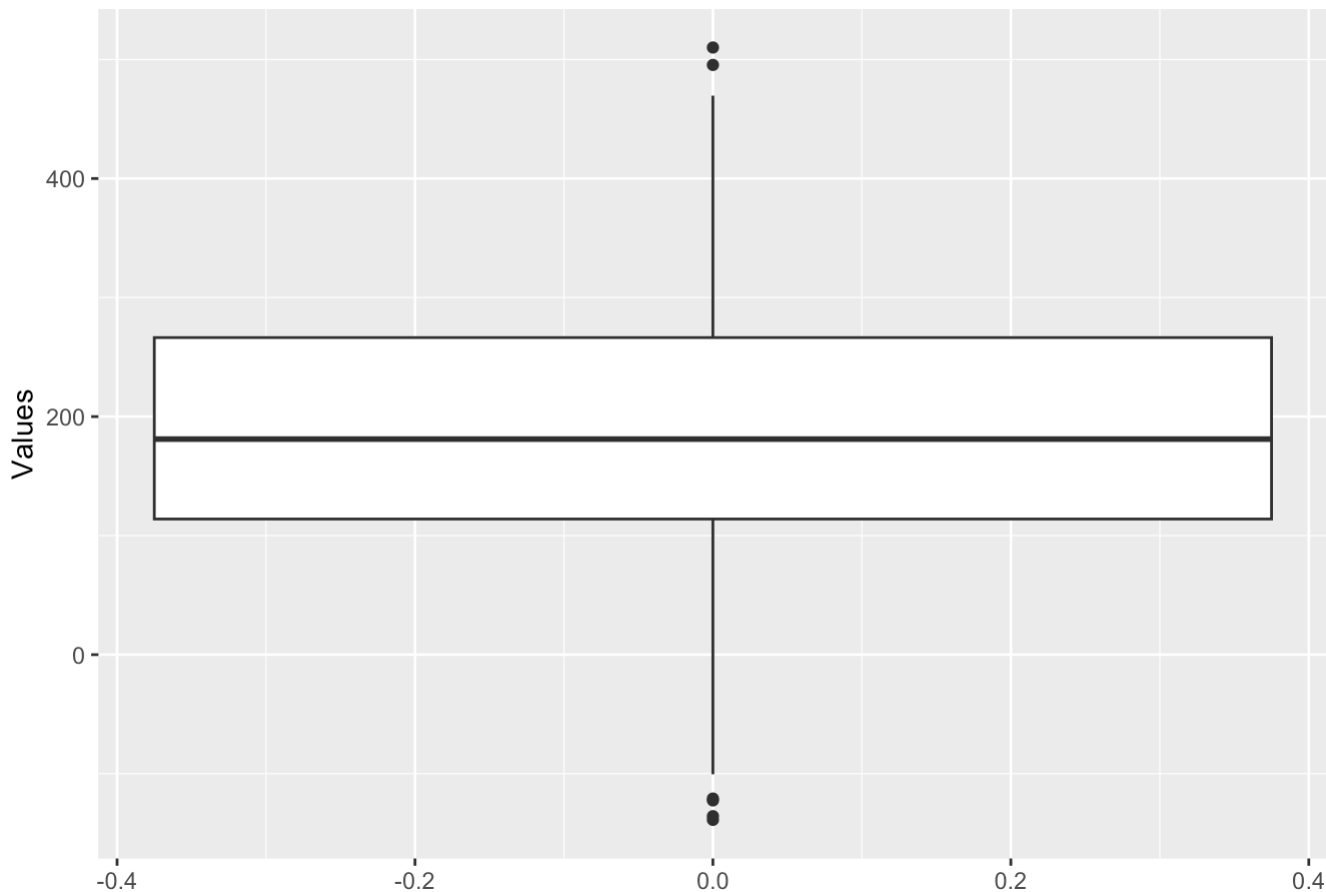
Boxplot of qb2\_game\_value



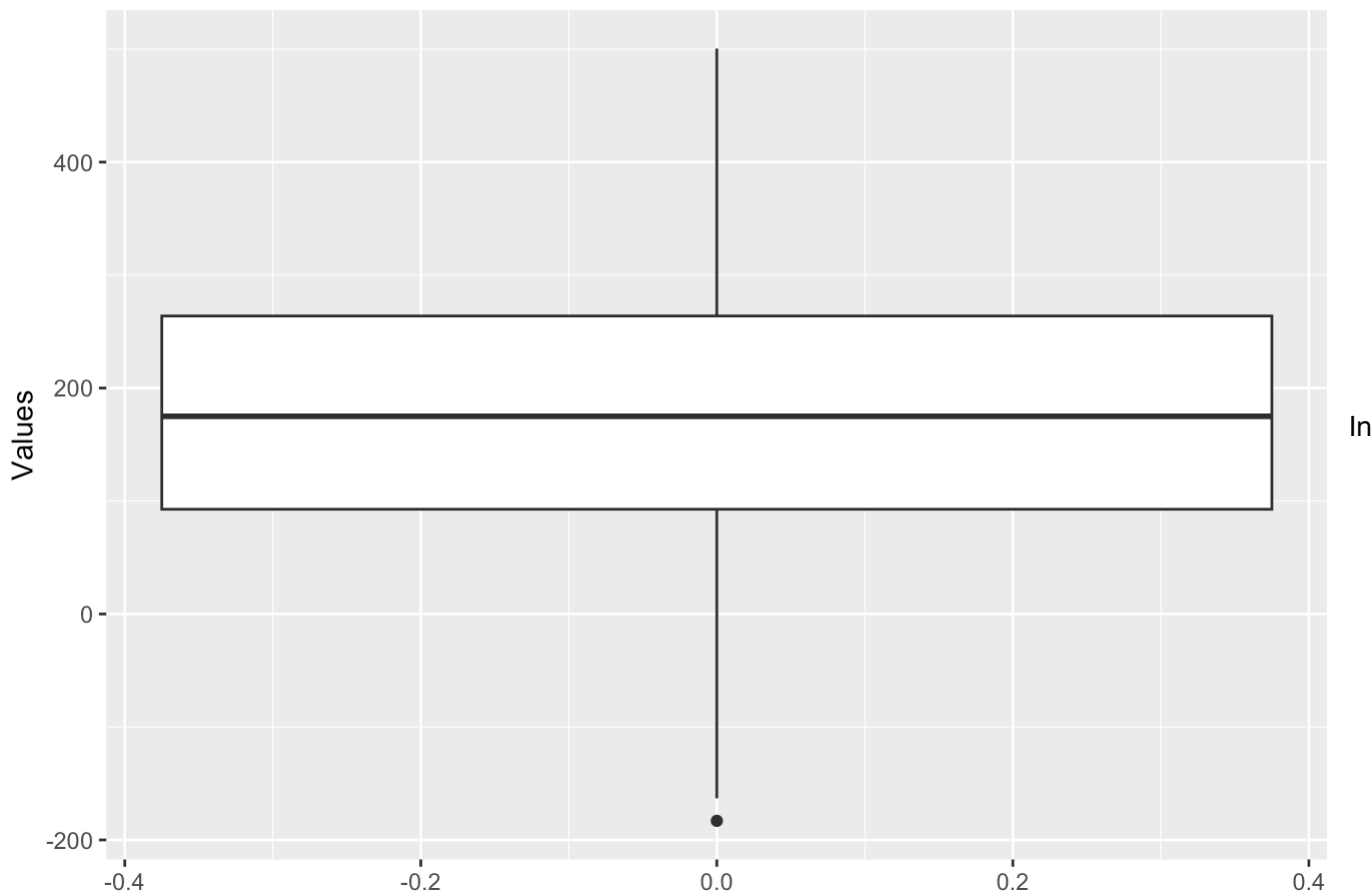
Both have positive and negative outliers. Let's call the `calculate_bounds` function to find the upper and lower bounds, then replace them with median

```
##  
## Summary for qb1_game_value :  
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
## -138.8   113.9   181.1   184.1   266.4   510.1  
##  
## Summary for qb2_game_value :  
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
## -183.10   92.73  174.96  176.33  263.80  500.55
```

Boxplot of qb1\_game\_value



Boxplot of qb2\_game\_value



prediction models, especially for sports like NFL, outliers may represent extreme scenarios, such as unusually bad or good performances predicted for a quarterback. These could be rare but realistic outcomes. If the outliers are natural and don't heavily impact your analysis, it's perfectly acceptable to leave them in the dataset.



## 17. Handling missing values in columns qb1\_game\_value, qb2\_game\_value, qbelo1\_post, and qbelo2\_post

```
##
## Summary for qb1_value_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
## -13.57 131.24 175.25 171.51 216.35 310.13     5
##
## Summary for qb2_value_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##   5.574 128.308 173.248 169.674 211.612 313.828     5
##
## Summary for qbelo1_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##    15    1437    1515    1507    1584    1757     5
##
## Summary for qbelo2_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##  1255    1444    1511    1513    1582    1755     5
```

qb1\_value\_post - negative outlier qbelo1\_post - outlier around 0

```
##
## Summary for qb1_value_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.431 133.957 175.249 175.125 215.067 310.131
##
## Summary for qb2_value_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.574 128.778 173.248 169.741 210.006 313.828
##
## Summary for qbelo1_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  1259    1441    1515    1513    1583    1757
##
## Summary for qbelo2_post :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  1255    1444    1511    1513    1580    1755
```

## 18. Handling missing values in columns score 1 and score 2

```
##
## Summary for score1 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
## -16.00   19.00   24.00   24.67   31.00   99.00     4
##
## Summary for score2 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##    0.00   17.00   25.00   24.62   31.00   49.00     4
```

Score 1 - Negative scores are considered outliers as they are unrealistic, clearly errors. Also, a score of 99 is extremely unlikely

```
##
## Summary for score1 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.00   19.50   24.00   24.65   31.00   56.00
##
## Summary for score2 :
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.00   17.00   25.00   24.63   31.00   49.00
```

```
## The cleaned dataset has been saved as: cleaned_dataset.csv
```